Review of NSF OCI EAGER and NSF OCI SDCI projects

Jie Li, Zhengyang Liu, and Malathi Veeraraghavan
University of Virginia
{jl3yh, zl4ef, mvee}@virginia.edu
March 22, 2012

This work was carried out as part of NSF sponsored research projects, OCI-1038058 and OCI-1127340

Agenda

- SDCI Project Review
  - Data analysis of scientists' file transfer logs
  - 100 Gbps testing from NERSC to ANL on ANI testbed
  - Our ANI 100G testbed experiments
  - Ongoing work
- EAGER Project Review
  - Analysis and selection of a network service for the UCAR scientific data distribution project
  - Design and implementation of the Virtual Circuit Multicast Transport Protocol (VCMTP)
- DYNES participation

Publications

- J. Li, M. Veeraraghavan, M. Manley and S. Emmerson, "Analysis and selection of a network service for a scientific data distribution project," Proc. IEEE CMC 2012, May 2012
- J. Li, M. Veeraraghavan, "A Reliable Message Multicast Transport Protocol for Virtual Circuits," Proc. IEEE CMC 2012, May 2012
- Z. Liu, M. Veeraraghavan, Z. Yan, C. Tracy, J. Tie, I. Foster and J. Dennis, "Science traffic characterization and network service selection," submitted to IEEE HPSR 2012
- Z. Yan, C. Tracy, M. Veeraraghavan, "A Hybrid Network Traffic Engineering System," submitted to IEEE HPSR 2012

GridFTP transfers

- Analyzed GridFTP usage statistics to answer two questions:
  - Are the high-throughput file transfer sessions long enough to justify VC setup delay (current number: 1 min)?
    - NCAR and SLAC data analysis
    - Use 3rd quartile throughput and 10 mins for duration
  - Is throughput variance caused by competing IP-routed network traffic?
    - If so, VCs useful to guarantee rate for science flow
    - NERSC data analysis

GridFTP usage stats

- For each transfer, servers log size, start time, duration, number of parallel streams, stripes, dest IP
- Usage stats are sent via UDP to Globus server from site GridFTP server
- Stats can be obtained with permission from Globus or the site itself (e.g., NCAR, SLAC, NERSC, BNL)

5

Find "sessions" from transfers

- Typical scientist uses shell scripts to move "Lots of Small Files (LOSF)"
- From GridFTP usage stats, need an algorithm to "merge" transfers into sessions
- Multiple simultaneous transfers; look for last completion time
- If next transfer's start time is within 1 minute of completion time, we assume it is part of the same session

6

NCAR-NICS GridFTP data

Sessions	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Actual size (MB)	0	5256	69800	256500	318900	2607000
Actual durations (sec)	0.05	188.9	1445	4029	5250	48420

Transfers	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Throughput (Mbps)	0	296.9	468	505.5	681.7	4227

- 2009-2011, but only two users
- max-duration session was 13 hrs 27 mins (48420 s) but size was only 2.4 TB (rate: 410 Mbps)
- max-size session (2.7 TB) took 7.5 hours (808 Mbps)

7

Thanks to John Dennis and Matt Woitaszek, NCAR

SLAC-BNL GridFTP usage logs (≥100MB)

Sessions	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Actual size (MB)	104	633	1734	17430	5702	3595000
Actual durations (sec)	2.03	29.7	77.8	282.8	172.1	35820

Transfers	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Throughput (Mbps)	0.013	25.12	127	136.3	191.3	1930

- Number of transfers much larger than NCAR-NICS
- 3rd quartile throughput much smaller
- Note that the session that is "max" from a size perspective is not necessarily the one that is "max" from a duration perspective

8

Thanks to Yee Ting Li and Wei Yang, SLAC

% of sessions for which dynamic VCs are suitable

- NCAR-NICS (2009-2011)
  - 217 sessions from 52519 transfers
  - 197 sessions >= 100MB
  - 63% of >=100MB sessions would > 10mins if they experienced third quartile throughput of 681.7 Mbps
  - longest session: 13.5 hrs; size: 2.4TB (410Mbps)
  - max size session: 2.7 TB; dur = 7.5 hours (808Mbps)
- SLAC-BNL (Feb. 10-24, 2012)
  - Throughput (3rd quartile: 191 Mbps; max: 1.93 Gbps)
  - 2233 sessions from 133,346 transfers
  - 1977 sessions >= 100MB
  - 13.4% of sessions would have lasted longer than 10 mins if they had experienced a throughput of 191 Mbps

9

NERSC data

- GridFTP transfers from NERSC DTN servers that > 100 MB in one month (Sept. 2010)
- Total number of transfers: 124236
- GridFTP usage statistics

TABLE I: Summary of all NERSC transfers larger than 100 MB; the three columns are independent, e.g., the transfer with the largest size is not the same transfer as the one with the longest duration or the one with the highest throughput

	Size (Bytes)	Duration (s)	Throughput (bps)
Min	1.000e+08	0.2488	1.266e+06
1st Quartile	1.049e+08	1.9229	1.713e+08
Median	1.049e+08	2.4919	3.480e+08
Mean	2.531e+08	35.4022	3.557e+08
3rd Quartile	1.261e+08	8.8897	4.445e+08
Max	9.679e+10	9952.2382	4.315e+09

Thanks to Brent Draney, Jing Tie and Ian Foster for the GridFTP data

10

NERSC data session analysis

- Obtained NERSC data from Globus (Ian Foster)
- The usage stats reported to Globus does not include dest IP address (privacy reasons)
- Cannot group transfers into sessions
- Working with Brent Draney and Jason Hick, NERSC, to have them assign someone to run our analysis code

11

Usage of dynamic VCs

- Some percentage of sessions are long-lived even if rate of the transfers is assumed to be high
- Therefore dynamic VCs can be setup
- Ideally, VC setup delay should be reduced
- Dynamic VCs important for inter-domain science flows
  - ESnet – Hurricane Electric experience

12

GridFTP transfers

- Analyzed GridFTP usage statistics to answer two questions:
  - Are the high-throughput file transfer sessions long enough to justify VC setup delay (current number: 1 min)?
    - SLAC and NCAR data analysis
    - Use 3rd quartile throughput and 10 mins for duration
  - ➢ Is throughput variance caused by competing IP-routed network traffic?
    - NERSC data analysis

Throughput variance

TABLE II: Summary of all 32 GB NERSC transfers

	Duration (s)	Throughput (bps)
Min	75.4	7.579e+08
1st Qu.	141.20	1.251e+09
Median	183.40	1.499e+09
Mean	186.60	1.625e+09
3rd Qu.	219.70	1.947e+09
Max	362.70	3.644e+09

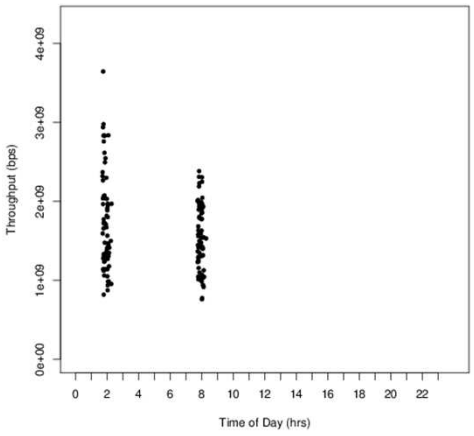
- There were 145 file transfers of size 32 GB to ORNL
  - Same round-trip time (RTT), bottleneck link rate and packet loss rate
- IQR (Inter-quartile range) measure of variance is 695 Mbps
- Find an explanation for this variance

Potential causes of throughput variance

Same for 145 transfers

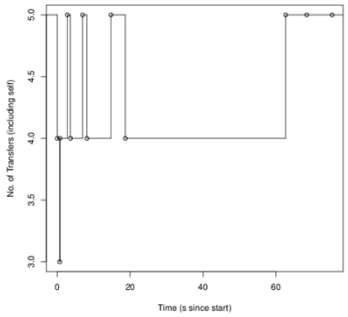
- Path characteristics:
  - RTT, bottleneck link rate, packet loss rate
- Number of stripes
- Number of parallel TCP streams
- Time-of-day dependence
- Concurrent GridFTP transfers
- Network link utilization (SNMP data)
- CPU usage, I/O usage on servers at the two ends

Time-of-day dependence (NERSC 32 GB: same path)



- Two sets of transfers: 2 AM and 8 AM
- Higher throughput levels on some 2 AM transfers
- But variance even among same time-of-day flows

Dep. on concurrent transfers: Predicted throughput

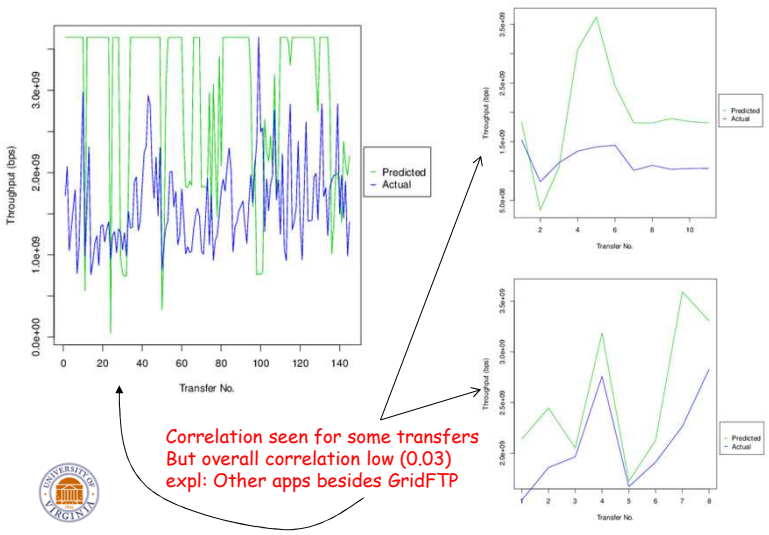


$$\tilde{T}_i = T_{max} \sum_{j=1}^{j_{max}} \frac{1}{n_{ij}} \times \frac{d_{ij}}{D_i}$$

- Find number of concurrent transfers from GridFTP logs for i^{th} 32 GB GridFTP transfer: NERSC end only
- Determine predicted throughput
- d_{ij}: duration of j^{th} interval of i^{th} transfer
- n_{ij}: number of concurrent transfers in j^{th} interval of ith transfer

17

Dependence on concurrent transfers (NERSC 32 GB transfers)



18

Correlation with SNMP data

Correlation between GridFTP bytes and total SNMP reported bytes

	if1	if2	if3	if4	if5
1st Qu.	0.677	0.604	0.719	0.750	0.749
2nd Qu.	0.419	0.147	0.138	0.327	0.294
3rd Qu.	0.538	0.592	0.543	0.415	0.371
4th Qu.	0.782	0.872	0.797	0.789	0.790
All	0.902	0.922	0.919	0.918	0.918

Correlation between GridFTP bytes and other flow bytes

	if1	if2	if3	if4	if5
1st Qu.	0.254	0.188	0.429	0.505	0.486
2nd Qu.	0.269	-0.067	-0.110	0.089	0.071
3rd Qu.	0.059	0.157	0.110	0.015	-0.039
4th Qu.	0.196	0.328	0.239	0.287	0.276
All	0.351	0.365	0.443	0.524	0.527

- Got SNMP data for ESnet links on NERSC-ORNL path
- SNMP raw byte counts: 30 sec polling
- Assume GridFTP bytes uniformly distributed over duration
- Conclusion: GridFTP bytes dominate and are not affected by other transfers – consistent with alpha behavior
- Use of VCs may not solve throughput variance problem

Thanks to Jon Dugan for the SNMP data

Still pending for this variance study

- For the NERSC-ORNL transfers
  - Need SNMP data for links inside NERSC and inside ORNL
  - Need CPU and I/O usage data at the two servers
  - Common belief: cause of variance is file system access
  - Computing nodes write to file systems while DTNs read file systems
  - Working with Brent Draney and Jason Hick, NERSC, and Galen Shipman, ORNL, for site data

NCAR-NICS throughput variance

TABLE X: Throughput of 16GB/4GB transfers in NCAR data set(Unit: Mbps)

Year based analysis of 16GB transfers								
Year	No. of Transfers	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Standard Deviation
2009	1076	10.79	707.3	889.3	877	1075	1543	294.17
2010	233	95.36	516.5	619.2	651.7	742.9	1150	205.53
2011	12	441.73	480.71	538.77	539.1	575.39	652.07	66.92
Year based analysis of 4GB transfers								
Year	No. of Transfers	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Standard Deviation
2009	853	4.14	593.1	873.1	849.2	1125	1587	366.01
2010	247	72.99	767	977.1	903.1	1083	1209	225.09
2011	37	296.27	376.13	497.81	475.63	526.6	637.13	101.12

TABLE XI: Throughput of 16GB/4GB transfers in NCAR data set(Unit: Mbps)

Stripes based analysis of 16GB transfers								
No. of Stripes	No. of Transfers	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Standard Deviation
1	13	441.73	483.7	541.75	546.84	616.48	652.07	69.88
2	547	10.79	542	714	705.4	855.2	1207	212.34
3	761	19.83	748.7	976	931.6	1150	1543	306.96
Stripes based analysis of 4GB transfers								
No. of Stripes	No. of Transfers	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Standard Deviation
1	18	372.2	449.6	506.2	569.2	574.2	1309	225.85
2	447	72.99	566.7	773.1	772.8	1021	1209	245.37
3	759	4.14	625.6	927.6	875.8	1169	1587	375.38

- Clear dependence on number of stripes
- NCAR reduced number of servers from 3 to 1 in 2009-2011 period

21

Next steps

- Run a set of controlled experiments on ANI testbed and experiment with tools for obtaining CPU usage and disk I/O (file system) usage measurements for regression analysis with GridFTP transfer throughput
- Instrument servers at sites and collect data to explain causes of variance
  - NERSC-ORNL: Jason Hick and Galen Shipman
  - SLAC-BNL: Yee Ting Li and Scott Bradley
  - NCAR-NICS: John Dennis and Victor Hazelwood

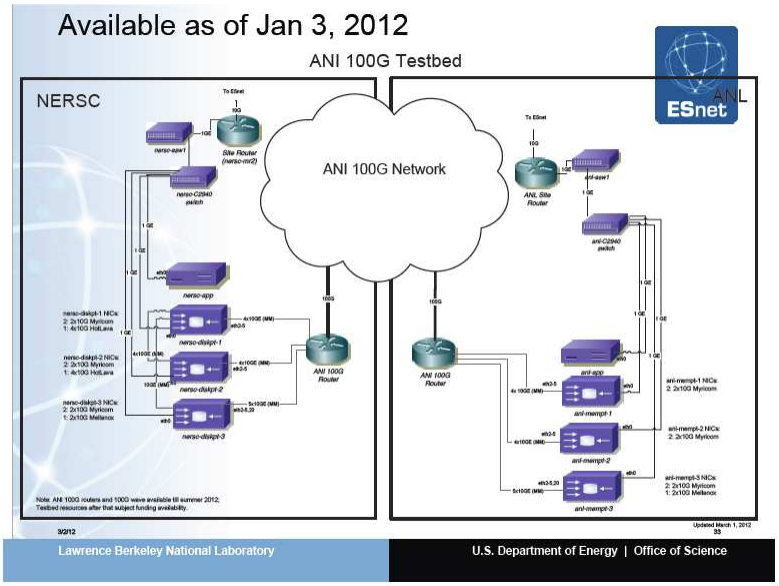
22

Agenda

- SDCI Project Review
  - Data analysis of CESM scientists' logs
  - ➢ 100 Gbps testing from NERSC to ANL on ANI testbed
  - ➢ Our ANI 100G testbed experiments
  - Ongoing work
- EAGER Project Review
  - Analysis and selection of a network service for the UCAR scientific data distribution project
  - Design and implementation of the Virtual Circuit Multicast Transport Protocol (VCMTP)
- DYNES participation

ANI 100G Testbed

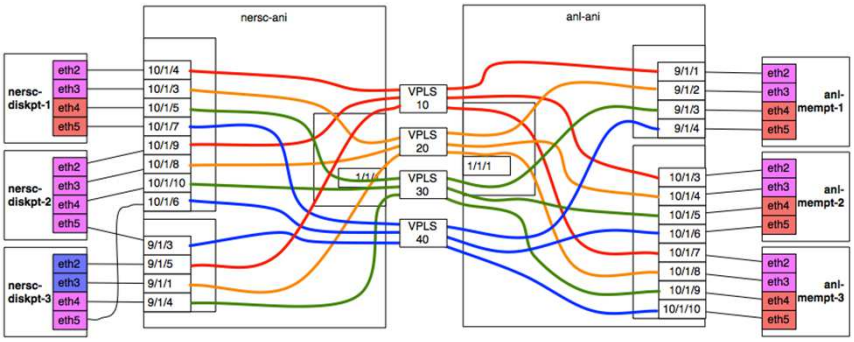
Available as of Jan 3, 2012



Brian Tierney DOE PI meeting, March 1-2, 2012

ANI 100G Testbed Experiments

- Performance: 48.6ms RTT, 97.9Gbps aggregate TCP throughput with 10 TCP streams



Brian Tierney's DOE PI meeting talk, March 1-2, 2012
Work done by Eric Pouyol and Brian Tierney, ESnet

ANI 100G Testbed Experiments

```
nersc-diskpt-1-v4012:  1179.1875 MB /   1.00 sec = 9891.8010 Mbps     0 retrans
nersc-diskpt-1-v4013:  1179.2500 MB /   1.00 sec = 9888.4787 Mbps     0 retrans
nersc-diskpt-1-v4014:  1179.1875 MB /   1.00 sec = 9891.1482 Mbps     0 retrans
nersc-diskpt-1-v4015:  1179.1250 MB /   1.00 sec = 9891.1581 Mbps     0 retrans
nersc-diskpt-2-v4012:  1179.2500 MB /   1.00 sec = 9891.9494 Mbps     0 retrans
nersc-diskpt-2-v4013:  1179.0625 MB /   1.00 sec = 9891.1580 Mbps     0 retrans
nersc-diskpt-2-v4014:  1179.3750 MB /   1.00 sec = 9893.1365 Mbps     0 retrans
nersc-diskpt-2-v4015:  1179.1250 MB /   1.00 sec = 9891.0690 Mbps     0 retrans
nersc-diskpt-3-v4014:  1121.8750 MB /   1.00 sec = 9410.9602 Mbps     0 retrans
nersc-diskpt-3-v4015:  1121.8750 MB /   1.00 sec = 9410.9884 Mbps     0 retrans

                                                  Input                Output
-------------------------------------------------------------------------------
Octets                                         18462079          12387383345
Packets                                          184615              1369129
Errors                                                0                    0
Utilization (% of port capacity)                   0.17                99.31
```

Brian Tierney's DOE PI meeting talk, March 1-2, 2012

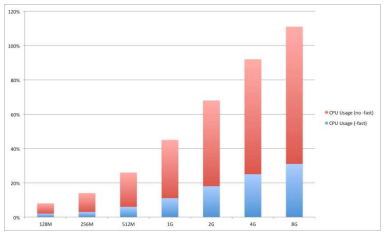
UVA's ANI testbed experiments to date

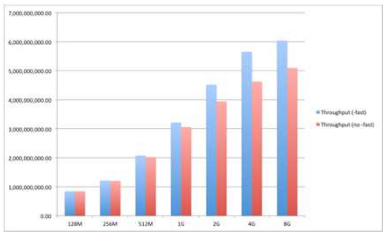
- GridFTP, iperf and nuttcp transfers between NERSC and ANL (up to 30 Gbps; difficult to reserve whole testbed)
- CPU usage becomes the limiting factor of throughput under high bandwidth
  - GridFTP client utilizes 100% CPU when throughput is 5.4Gbps; need second core
  - iperf and nuttcp: 34% CPU for 9.4 Gbps

27

GridFTP fast option testing data size: 128 MB to 8GB

- The "-fast" option of GridFTP relieves pressure on CPU on client side (still reaches 100% on server side when throughput reaches 9.4Gbps)
- Conclusion: need to experiment with RNICs and verbs interface (TCP/IP in O/S consumes CPU cycles)





CPU Usage (memory-to-memory transfer, NERSC-ANL)

Throughput (memory-to-memory transfer, NERSC-ANL)

ANI 100G testbed experiments

- Planned for March 23, 2012:
  - RoCE across WAN: Bob Russell's programs for latency, throughput, CPU utilization
  - GridFTP with UDT
- Next steps:
  - Add verbs interface module to GridFTP (UNH)
  - Test GridFTP across RoCE (nersc-diskpt3 to anl-mempt3) with wide-area VCs

Related work

- EXS API: UNH (Russell)
- CCI: ORNL (Atchley, Shipman)
- ADTS: Ohio State Univ (DK Panda)
- XSP: Delaware/IU (Kissel/Swany)
- UDT and TCP/IP: IPoIB and SDP

Intra-datacenter work

- Use Carver or Lawrencium, NERSC, and Cray (kraken)
  - IB clusters, and Seastar, Gemini interconnects
- UNH will develop plan for data collection, and instrument
- NCAR will run CESM apps and benchmarks
- UVA will analyze data
- UNH is developing course modules

Agenda

- SDCI Project Review
  - Data analysis of scientists' file transfer logs
  - 100 Gbps testing from NERSC to ANL on ANI testbed
  - Our ANI 100G testbed experiments
  - Ongoing work
- ➢ EAGER Project Review
  - Analysis and selection of a network service for the UCAR scientific data distribution project
  - Design and implementation of the Virtual Circuit Multicast Transport Protocol (VCMTP)
- DYNES participation

EAGER project motivation

- Large scale scientific data sets are increasingly distributed to geographically dispersed research organizations/scientists
- Different types of network services
  - IP-routed service vs. Virtual circuits
  - Unicast vs. Multicast
  - P2P
- Problem statement
  - What is the best network service for scientific data distribution?

33

Background

- IP-routed service
  - ubiquitous
  - offers reliable data delivery using TCP
- Static circuit service
  - Offers a dedicated circuit between two or more endpoints for a pre-specified duration
- Dynamic circuit service (DCS)
  - Connect to any other DCS subscriber for rate-guaranteed communications for specified durations
- Research-and-education network (REN) and commercial providers now offer dynamic circuit service

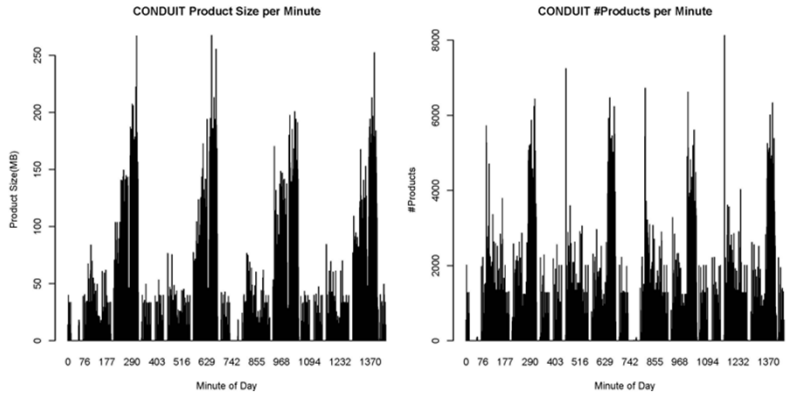
34

Case Study

- Internet Data Distribution (IDD)
  - Meteorology data distribution project run by the University Corporation for Atmospheric Research (UCAR)
  - Near real-time data distribution system to over 160 institutions
  - Software called Local Data Manager (LDM) is used for data distribution
  - Over 30 types of scientific data products (feedtypes) are distributed using LDM

35

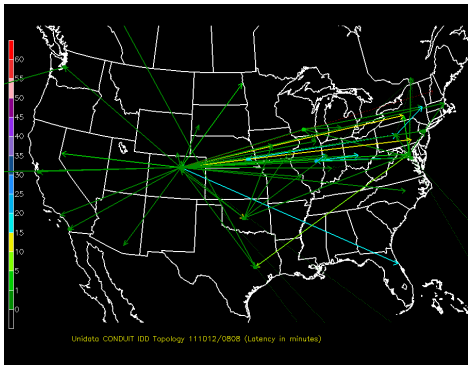
Analysis of the CONDUIT Feedtype



- Total size per day: ~60 GB
- Peak throughput: 250 MB per minute (33.3 Mbps)
- Less than 2% of silence periods are larger than 1 second

36

CONDUIT Distribution Topology

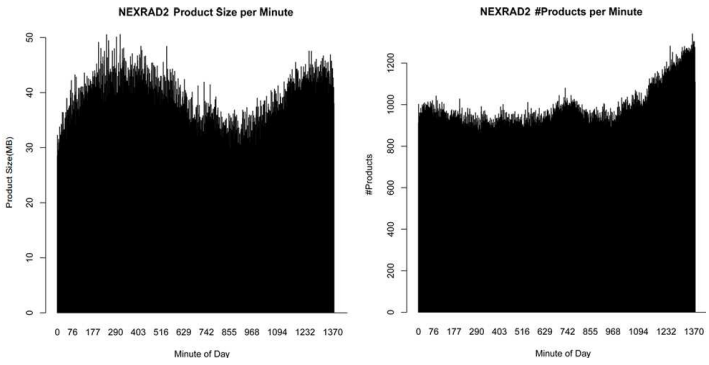


Parameter	Number
Total number of Distinct Hosts	163
# Sender Hosts	57
# Receiver Hosts	141
Max. Fan-out Number	104

- 104 receivers are directly connected to the UCAR IDD servers (the maximum fan-out number)
- Bandwidth requirement: 104 * 33.3 Mbps = 3.5 Gbps

37

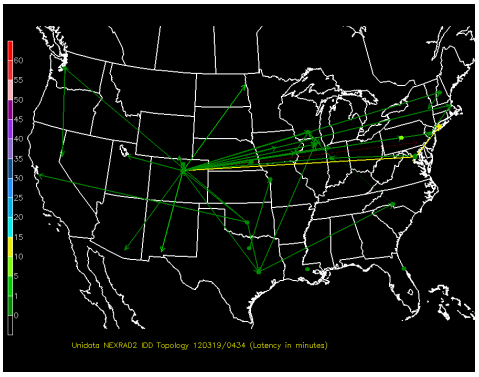
Analysis of the NEXRAD2 Feedtype



- Total size per day: ~56 GB
- Peak throughput: 58 MB/minute (7.8 Mbps)
- Almost all silence periods are less than 1 second

38

NEXRAD2 Distribution Topology



Parameter	Number
Total number of Distinct Hosts	150
# Sender Hosts	75
# Receiver Hosts	114
Max. Fan-out Number	55

- IDD servers at UCAR directly deliver NEXRAD2 data to 55 receivers
- Bandwidth requirement: 55 * 7.8 Mbps = 429 Mbps

39

Selection of A Suitable Network Service

- Current network service used by IDD
  - Unicast TCP connections over IP-routed paths
  - Data products are effectively sent to the receivers in a round-robin fashion
  - Pros: service is ubiquitous
  - Cons: requires UCAR to run 9 servers for IDD; uses 5 Gbps of its access link; data delivery latency sensitive to the number of receivers

40

Selection of A Suitable Network Service (cont.)

- Static unicast virtual circuits?
  - May be good for NEXRAD2, but bad for CONDUIT due to its burstiness
  - Utilization will be poor if circuit rates are chosen to be high to keep latency low
- Dynamic circuit service?
  - DCS can be scheduled for the CONDUIT bursty periods
  - BUT the silence periods are too short (mostly less than 1 second) for circuits to be scheduled and set up for use (setup delay ~1 min in today's REN offerings)

Selection of A Suitable Network Service (cont.)

- Multicast virtual circuits
  - Unlike IP multicast, no potential data-plane congestion in rate-guaranteed virtual circuits
  - Negative acknowledgements (NACKs) used
  - Packet loss due to receive buffer overflows or bit errors will be handled at the end of the multicast
    - important for high-speed multicast
  - Our hypothesis: the throughput for most receivers in a VC multicast group can be independent of the throughput experienced by some slow receivers that incur retransmissions

P2P vs. multicast

- P2P requires more than one transfer for most of the blocks
- Multicast requires one transfer + retransmissions for lost blocks (small percent with real-time scheduling)
- P2P suitable if file is already available in multiple nodes, but in IDD, files are available only at a single node in the beginning and needs to be distributed quickly before next arrival

Agenda

- SDCI Project Review
  - Data analysis of scientists' file transfer logs
  - 100 Gbps testing from NERSC to ANL on ANI testbed
  - Our ANI 100G testbed experiments
  - Ongoing work
- EAGER Project Review
  - Analysis and selection of a network service for the UCAR scientific data distribution project
  - ➢ Design and implementation of the Virtual Circuit Multicast Transport Protocol (VCMTP)
- DYNES participation

Requirements for VCMTP

- Reliability
  - Error control, flow control
- Scalability
  - One multicast group should support hundreds of receivers
- Design goal: one slow receiver that incurs retransmissions will not decrease the throughput for all receivers

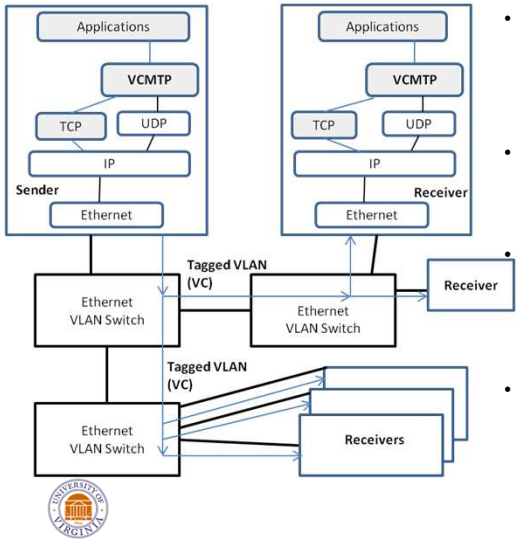
45

VCMTP Key Design Concepts

- For high-speed transfers, multicast whole file before handling retransmissions
  - future version: relax to allow fast senders or fast receivers to run retransmission thread in parallel
- Run VCMTP sender/receiver processes in high-priority mode (SCHED_RR)
  - Decreases receive buffer overflow losses
- Unicast TCP connections for retransmissions
- Negative Acknowledgement (NACK) to avoid positive ACK-implosion problem
- Multicast groups with different send rates serve different groups of receivers

46

VCMTP Prototype



- Data blocks of a message are encapsulated in UDP packets to be multicast over Ethernet
- Blocks written to disk using offset (out of sequence writes)
- Receivers send retransmission requests to the sender over unicast TCP connections
- Sender has multiple retransmission threads each with a unicast TCP connection to a receiver

47

Experimental Testbed

- Emulab Testbed
  - Located at the University of Utah
  - Over *500* nodes (both high-end and low-end) connected by high-end switches and routers
  - High-end nodes (D710 series): 2.4 GHz 64-bit Quad Core Xeon E5530, 12 GB RAM, 1 Gbps links
  - Low-end nodes (PC600 series): 600MHz Intel Pentium III, 256 MB RAM, 100Mbps links

48

Evaluation of Multicast Performance

- One sender multicasts disk files of different sizes to 7 high-end receiver nodes (D710)
- VCMTP is the only user process running on the nodes
- Sending rate: 600 Mbps
- For each file size, the data multicast is repeated 10 runs (7 * 10 = 70 receptions)

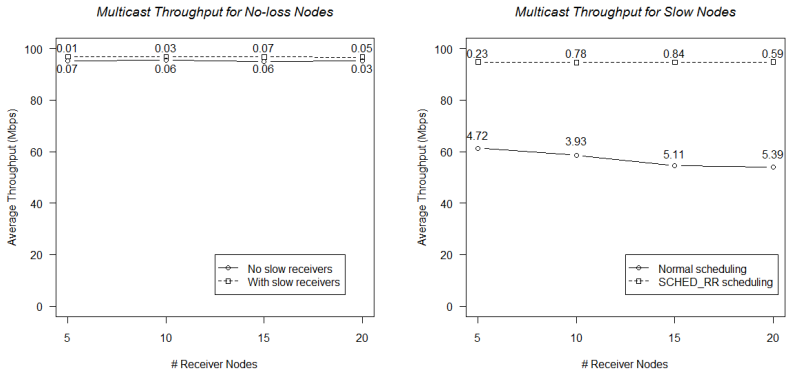
	512 MB	1 GB	2 GB	4 GB
Avg. (SD) throughput of receptions in no-loss runs	579.49 (1.73)	574.56 (1.60)	588.25 (0.30)	582.17 (0.87)
Avg. (SD) throughput of no-loss receptions in loss runs	N/A	575.65 (0.81)	588.27 (0.74)	582.22 (0.98)
Avg. (SD) throughput of loss receptions in loss runs	N/A	561.4 (1.73)	580.32 (4.94)	576.1 (4.43)

No degradation in throughput for fast receivers in the presence of slow receivers

Effects of multitasking and SCHED_RR scheduling

- When the VCMTP process is running with other processes on the receiver node, packet loss may occur due to resource sharing (CPU, I/O, etc.)
- Our solution: run the VCMTP process in higher priority than other processes
- Linux provides support for process scheduling with soft real-time priority (SCHED_RR mode)
- In this experiment, one sender multicasts a 128-MB disk file to X low-end receiver nodes (PC600), where 20% of the X nodes run the VCMTP process along with two other CPU-intensive benchmarks (*double* and *fstime* from the UnixBench suite)

Effects of multitasking and SCHED_RR scheduling (cont.)

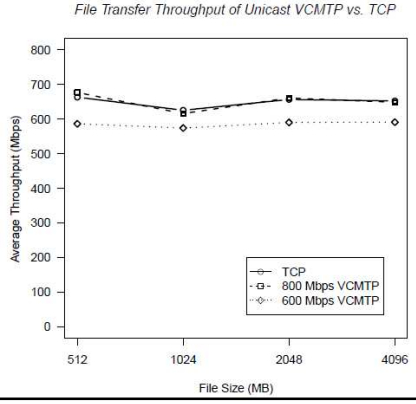


- 128 MB file multicast tests with 5, 10, 15, and 20 nodes
- Each file transfer was repeated 10 times for each expt.
- An expt: particular set of "slow" nodes (repeat 5 times)
- Standard deviations are shown as numbers in the plots

51

Negative of VCMTP: need to select sending rate

- File transfer experiments with single-sender, single-receiver for both TCP and VCMTP (on D710 nodes)
- Two different sending rates (800 Mbps and 600 Mbps)
- Each file transfer was repeated 10 times



Avg. Retransmission Rates for VCMTP

	600 Mbps	800 Mbps
512 MB	0%	4.46%
1 GB	0.26%	11.04%
2 GB	0.07%	8.63%
4 GB	0.19%	9.27%

52

Latency Analysis for TCP vs. VCMTP

- Consider the scenario where a sender sends a message of size s to n receivers
- Maximum throughput supported by the sender is min{c_s, r_s}, where c_s is the link capacity, and r_s is maximum sending rate with 100% resource usage (CPU, IO, etc.)
- Similarly, maximum throughput supported by a receiver is min{c_r, r_r}
- Any one of these four capacity limitations could be the bottleneck

I. Total delay for unicast TCP

$$t_{tcp} = max\{\frac{ns}{r_s}, \frac{ns}{c_s}, \frac{s}{r_r}, \frac{s}{c_r}\}$$

* When r_s is the bottleneck, the total delay for unicast TCP can be reduced by running multiple sender servers in parallel

II. Total delay for VCMTP

$$t_{vcmtp} = max\{\frac{s}{r_s}, \frac{s}{c_s}, \frac{s}{r_r}, \frac{s}{c_r}\}$$

Example

- Consider the case where s = 125 MB, n = 50, r_s = r_r = 1 Gbps, c_s = 10 Gbps, c_r = 100 Mbps

I. For unicast TCP, r_s is the bottleneck for the message distribution (although each receiver link capacity is only 100Mbps, the total throughput that can be supported to all receivers is 100 Mbps * 50 = 5 Gbps).Therefore, the total delay using unicast TCP is

125 MB * 8 * 50 / 1 Gbps = 50 sec

II. For VCMTP, the bottleneck for the message multicast is the receiver link capacity (c_r). Hence the total delay is

125 MB * 8 / 100 Mbps = 10 sec

To achieve the same total latency, 5 servers are needed at the sending side for unicast TCP (each sender can simultaneously send the message to 10 receivers)

Next steps for analysis

- Loss cases
  - With unicast TCP over IP-routed paths
    - losses at routers due to competing traffic
    - low loss rates due to overprovisioning
  - With VCMTP
    - priority scheduling of VCMTP process reduces receive buffer losses to low rates

VCMTP summary

- A reliable multicast transport protocol appears to be scalable if underlying network offers VC service
- High-speed transfers requires VCMTP processes to be run in high-priority mode if sender/receivers are multitasking
- VCMTP can both reduce bandwidth usage and the overall delay for some large-scale, long-duration data distribution tasks

DYNES

- Submitted proposal to Internet2 for UVA to become a DYNES end site
- Identified science users: LHC CMS physicist, Brad Cox, and biologist, Mike Timko at UVA – willing to try DYNES
- Use DYNES for VCMTP testing
  - Need logs in multiple nodes
  - Need to use NDDI OpenFlow for multipoint (ION for VCs)